CLAIMS

Having thus described our invention, what we claim as new and desire to secure by Letters Patent is as follows:

1. A method of controlling and guaranteeing a service level agreement (SLA) based on a communications outbound link bandwidth usage to a plurality of customers having electronic business activity hosted by at least one server as a server farm, said method comprising:

a) monitoring said outbound communications bandwidth usage by each customer traffic to determine a level of service being provided to each customer with respect to the agreed service level agreement in each service cycle time per unit of time;

b) controlling a flow of incoming requests to each customer business activity application so as to guarantee a level of service previously agreed to said customer by queuing and by selectively dropping requests to said customer to guarantee the agreed service levels to said customer,

wherein said controlling controls and guarantees each outbound link usage based service level agreement by controlling the flow of incoming requests to a server or group of servers.

2. The method as set forth in Claim 1, wherein said communications resource for service level agreements comprises an outbound communications link from said at least one server to said

network which sends responses from said server farm.

3. The method as set forth in Claim 1, wherein said communications resource customer traffic generated comprises a server farm having a plurality of servers, which is controlled by throttling the traffic to said server farm.

4. The method as set forth in Claim 1, wherein said communications resource customer traffic generated comprises a server farm, having at least one server, which is controlled by throttling the traffic to said server farm.

5. A method of controlling and guaranteeing the service level agreement (SLA) based on the communications outbound link bandwidth usage to a plurality of customers whose e-business and e-commerce are hosted by a server or a set of servers as a server farm, said method comprising:

a) monitoring said outbound communications bandwidth usage by each customer traffic to determine the level of service being provided to each customer with respect to the agreed service level agreement in each "service cycle time" per unit of time;

b) controlling the flow of incoming requests to each customer eBusiness/eCommerce application so as to guarantee a level of service previously agreed to said customer by queuing requests to said customer and by selectively dropping requests to said customer to guarantee the agreed service levels to said customer.

6. A method of guaranteeing a service level agreement (SLA) based on a communications outbound link bandwidth usage to a plurality of customers having electronic business activity hosted by at least one server, said method comprising:

a) monitoring said outbound communications bandwidth usage by each customer traffic to determine a level of service being provided to each customer with respect to the agreed service level agreement in each service cycle time per unit of time; and

b) controlling a flow of incoming requests to each said customer business activity application so as to guarantee a level of service previously agreed to said customer by queuing requests to said customer and by selectively dropping requests to said customer to guarantee the agreed service levels to said customer,

wherein said controlling controls and guarantees each outbound link usage based service level agreement by controlling the flow of incoming requests to the at least one server.

7. A method of regulating inbound requests on a world-wide network, comprising:

monitoring an amount of inbound traffic requests on a link of said world-wide network for a plurality of customers; and

regulating an output generated based on said amount of inbound traffic requests monitored in order to meet a service level agreement for said plurality of customers.

8. A system for controlling and guaranteeing a service level agreement (SLA) based on a communications outbound link bandwidth usage to a plurality of customers having electronic business activity hosted by at least one server as a server farm, said system comprising:

a) means for monitoring said outbound communications bandwidth usage by each customer traffic to determine a level of service being provided to each customer with respect to the agreed service level agreement in each service cycle time per unit of time; and

b) means for controlling a flow of incoming requests to each customer business activity application so as to guarantee a level of service previously agreed to said customer by queuing requests to said customer and by selectively dropping requests to said customer to guarantee the agreed service levels to said customer.

9. A communications system, comprising:

a worldwide network for communicating with a plurality of customers;

a manager, operatively coupled to said worldwide network, for controlling and guaranteeing a service level agreement (SLA) based on a communications outbound link bandwidth usage to said plurality of customers; and

at least one server functioning as a server farm, operatively coupled to said manager, said plurality of customers having electronic business activity hosted by said at least one server as said server farm.

10. The system according to claim 9, wherein said worldwide network comprises the Internet.

11. The system according to claim 10, wherein said customers output requests on said Internet for web data located on a cluster of servers.

12. The system according to claim 9, wherein said manager comprises a Communications Bandwidth Manager (CBM) presenting a single address to the at least one server cluster.

13. The system according to claim 12, wherein said manager includes a set of queues for queuing incoming requests.

14. The system according to claim 13, wherein said manager selects a request from the queues, selects one of the at least one server to service the request, and sends the request to that server.

15. The system according to claim 14, wherein said selected server receives the request, services said request, and sends a response directly back to said customer along a data output path,

wherein a portion of the path between said at least one server and said worldwide network is shared by a cluster of servers.

16. The system according to claim 15, wherein said manager controls an allocation of said outgoing data path among multiple customer sites hosted on the server cluster, by controlling incoming requests at the manager.

17. The system according to claim 9, wherein said manager comprises:

an input link for receiving incoming requests from customers; and

a set of queues for queuing the incoming requests.

18. The system according to claim 17, wherein said incoming requests are queued in order of arrival time and are serviced first-in, first out (FIFO) within each predetermined traffic class of requests.

19. The system according to claim 18, wherein said queues are provided for each traffic class for each customer.

20. The system according to claim 19, wherein the incoming requests include acknowledgment packets contain information on the quantity of outbound data that is being acknowledged, so as to estimate a volume of data that was output on an outgoing path from the server.

21. The system according to claim 19, wherein said manager monitors an outgoing data path to determine a number of data units delivered to a customer.

22. The system according to claim 21, wherein said manager, based on monitoring of said outgoing data path, determines if an amount of bandwidth being used by a customer exceeds a bandwidth amount per a service level agreement, such that feedback is generated to reduce the number of inbound requests being accepted.

23. The system according to claim 19, wherein said manager further comprises a traffic estimator for gathering monitored data on the usage of an output data path,

said traffic estimator also gathering output load information from said acknowledgment packets that arrive at the set of queues.

24. The system according to claim 23, further comprising a scheduler for receiving compiled load information from said traffic estimator,

wherein, based on the data from the traffic estimator, and service level agreement (SLA) information provided by an operator, said scheduler selects from requests in the queues, and for a selected request determines a server node to service the request, and sends the selected request to the server node.

25. The system according to claim 24, wherein said scheduler determines if requests or associated packets in the queues are to be discarded, or whether a direct response indicating an overloaded condition is to be sent to a customer.

26. A signal-bearing medium tangibly embodying a program of machine-readable instructions executable by a digital processing apparatus to perform a method of controlling and guaranteeing a service level agreement (SLA) based on a communications outbound link bandwidth usage to a plurality of customers having electronic business activity hosted by at least one server as a server farm, said method comprising:

a) monitoring said outbound communications bandwidth usage by each customer traffic to determine a level of service being provided to each customer with respect to the agreed service level agreement in each service cycle time per unit of time; and

b) controlling a flow of incoming requests to each customer business activity application so as to guarantee a level of service previously agreed to said customer by queuing requests to said customer and by selectively dropping requests to said customer to guarantee the agreed service levels to said customer.

27. A signal-bearing medium tangibly embodying a program of machine-readable instructions executable by a digital processing apparatus to perform a method of regulating inbound requests on a world-wide network, said method comprising:

monitoring an amount of inbound traffic requests on a link of said world-wide network for a plurality of customers; and

regulating an output generated based on said amount of inbound traffic requests monitored in order to meet a service level agreement for said plurality of customers.